Traffic engineering and GridFTP log analysis

Zhenzhen Yan, Z. Liu, M. Veeraraghavan
University of Virginia
mvee@virginia.edu

Chris Tracy, Chin Guok
ESnet
ctracy@es.net

Jan 17, 2013 Project web site:
http://www.ece.virginia.edu/mv/research/DOE09/index.html

Thanks to the US DOE ASCR for grants DE-SC002350 and DE-SC0007341 (UVA)
DOE for DE-AC02- 05CH11231 (ESnet)
NSF for grants, OCI-1127340 , OCI-1038058, and CNS-1116081 (UVA)

Thanks to Brian Tierney, Eric Pouyoul, Tareq Saif, Andy Lake & DOE: testbed

Thanks to Brent Draney, Jason Hick, NERSC, Yee-Ting Li, Wei Yang, SLAC, for GridFTP DTN login access

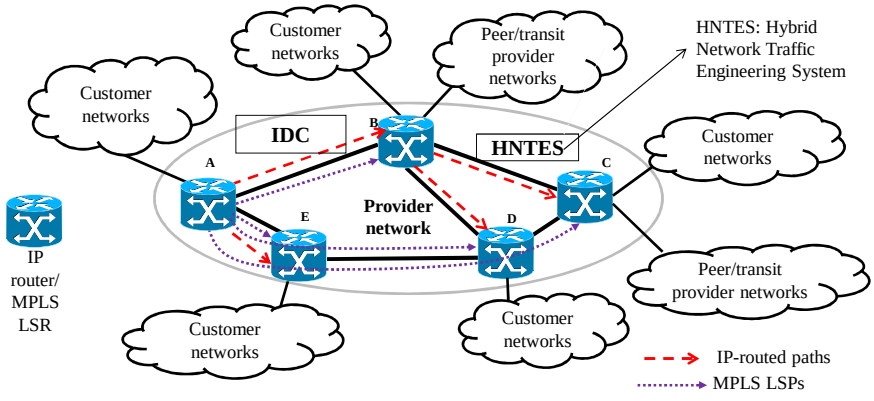
Outline

- Hybrid network traffic engineering system (HNTES)
  - alpha-flow identification and redirection
  - alpha flows: high-rate, large sized flows
  - threshold: 1 GB in 1 min
- GridFTP related work

Two key findings

- Throughput on circuits could be lower than on IP-routed paths
  - IDC circuit provisioning includes policing
  - Policing interacts with TCP in a negative way
  - But even with RoCE, throughput limited by circuit rate (typically < link capacity)
- Primary cause of throughput variance
  - CPU and I/O resources of DTNs
  - Application arguments (e.g., -fast)
  - Not too dependent on the net (overprovisioned)

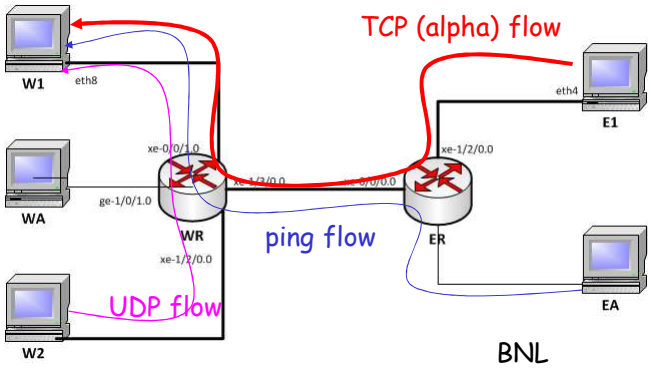
Hybrid network traffic engineering system (HNTES)
- Intradomain identification/redirection of alpha flows



- HNTES:
  - offline (e.g., nightly) analysis of NetFlow data at ingress routers
  - requests L3 circuits between ingress-egress router pairs from IDC
  - IDC sets firewall filters to direct alpha-flow packets to L3 circuits

Study router QoS configuration mechanisms

- Used DOE LIMAN testbed
- Hosts: high-performance diskpts



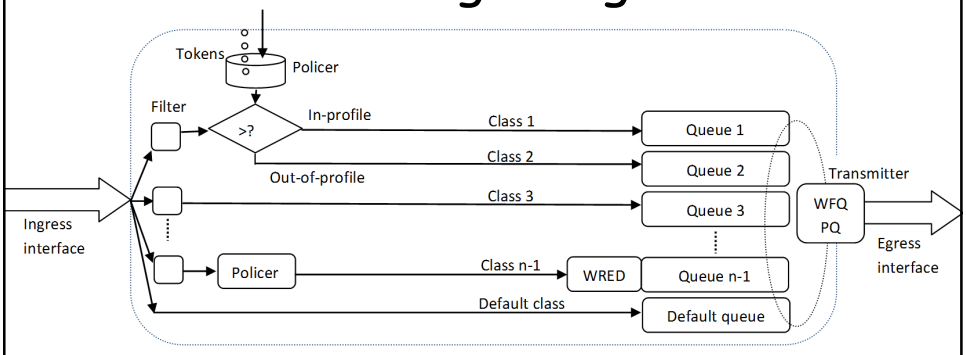
5

Router configurations

- Firewall filter:
  - alpha-flow based on /24 or /32 address prefixes (src and dst)
- Policing
  - classify out-of-profile packets as scavenger class and send to scavenger queue
  - Weighted Random Early Detection (WRED)
- Scheduling
  - Weighted fair queueing (WFQ)
  - Priority queueing (PQ)

6

Policing on ingress scheduling on egress



- Dual goals:
  - reduce impact of alpha flows on real-time delay-sensitive flows
  - allow alpha flows to enjoy high throughput

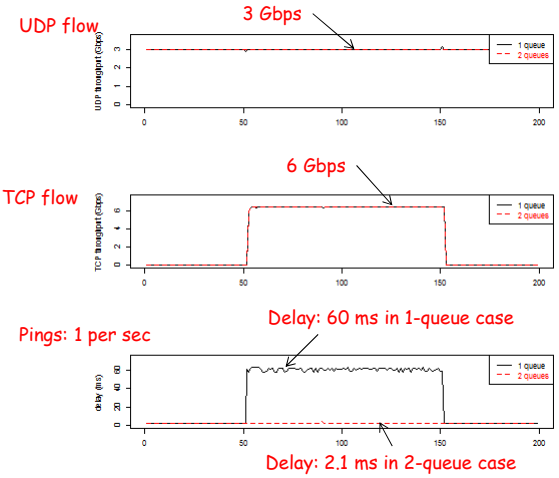
7

Compare 3 configurations

- 1-queue:
  - best-effort
  - all flows directed to same egress-side queue
- 2-queue: alpha and beta
  - scheduling-only (no policing)
    - WFQ +PQ
    - transmitter: shared in work conserving mode (non-strict)
    - buffer: strict partitioning
- 3-queue: alpha, beta, scavenger service (SS)
  - policing: > 1 Gbps sent to SS queue
  - scheduling: WFQ + PQ

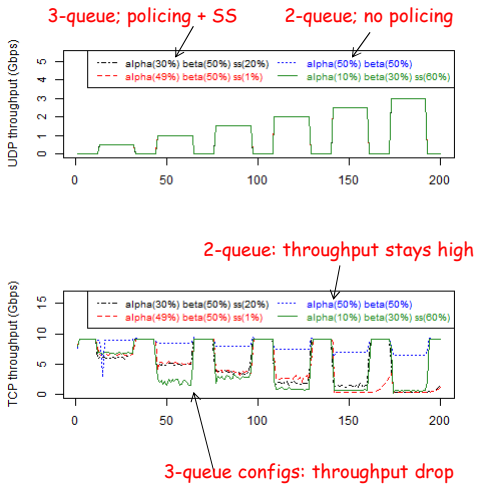
8

Impact of alpha flows on real-time flows



- Impact on ping flow delay
  - significant in 1-queue configuration
  - negligible in 2-queue configuration
- Need separate virtual queue for alpha flow packets

Impact of policing: 3-queue case causes TCP throughput to drop



- When UDP rate is increased from 0.5 to 1 Gbps, significant drop in TCP throughput
- Why? out-of-sequence packets
- TCP fast-retransmit/fast recovery algorithm causes sending rate to drop by half
- Worst when alpha queue allocation is only 10%

Key findings: HNTES

- Nodes generating alpha flows have static public IP addresses, and create repeated alpha flows
- Can leverage this fact in an offline HNTES design (nightly NetFlow analysis for alpha prefix ID determination)
- Configure ingress routers with firewall filters and WFQ/PQ 2-queue scheduling (alpha and beta); no policing
- ESnet5: WRED solution to policing
- IDC support requested: unspecified-rate circuits
  - implication: no policing

Z. Yan, M. Veeraraghavan, C. Tracy, C. Guok, "On how to provision Quality of Service (QoS) for large dataset transfers," submitted to CTRQ 2013

Outline

- Hybrid network traffic engineering system
- ➢ GridFTP
  - Usage log analysis software in R
  - GUI
  - Throughput variance:
    - competition for CPU and I/O within DTNs
    - less so for network resources

GridFTP usage log analysis & GUI

- Purpose: analyze and visualize GridFTP performance
- Obtained logs from
  - NERSC-ORNL (Sept. 2010)
  - NERSC-ANL (Mar. 4 - Apr. 22, 2012)
  - NCAR-NICS (2009-2011)
  - SLAC-BNL (Feb. 10 - Apr. 26, 2012)
- Analysis programs: coded in R
  - Session Analysis
  - Throughput Variance Analysis

Z. Liu, M. Veeraraghavan, Z. Yan, C. Tracy, J. Tie, I. Foster, J. Dennis, J. Hick, Y. Li, and W. Yang, "On using virtual circuits for GridFTP transfers," SC 2012.

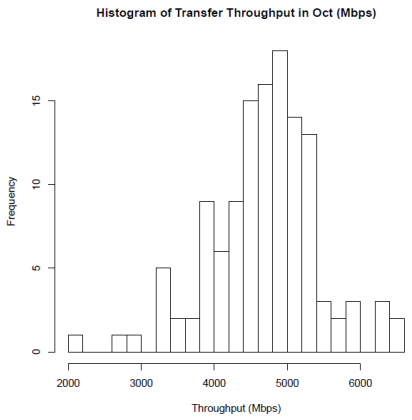
13

NERSC - SLAC Experiments To characterize causes of variance

- Path capacity: 10 Gbps
- At every hour
  - a memory to memory transfer runs from NERSC DTN to SLAC DTN for 30s
  - another one runs from SLAC to NERSC for 30s
  - prior to the two test runs, at the 59th minute every hour, the monitoring tools we developed are run to record CPU usage data and TCP traces
- SNMP data obtained from SLAC, NERSC, ESnet routers

14

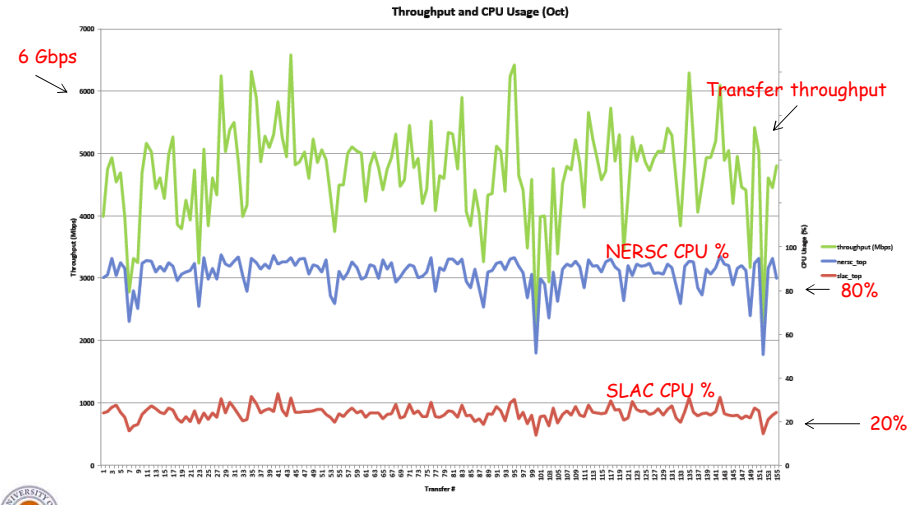
NERSC-SLAC GridFTP mem2mem transfer throughput



There is variance in transfer throughput

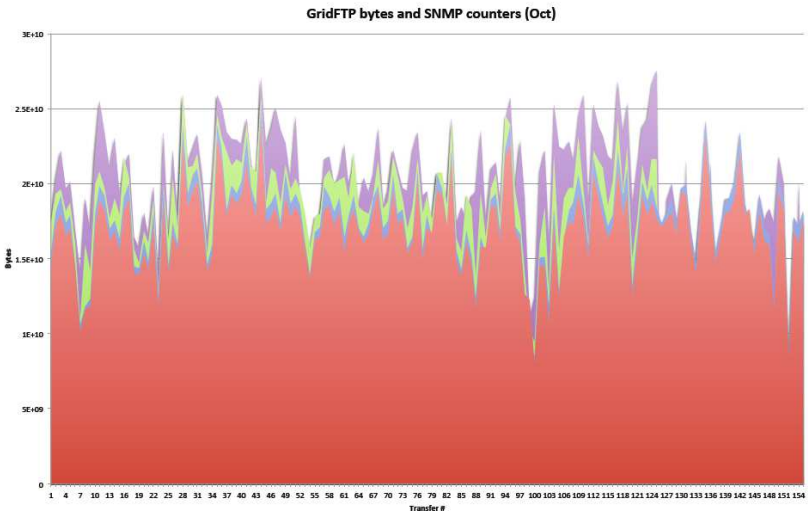
15

GridFTP transfer throughput dependence on CPU time



16

Network impact low



Correlation coefficients

GridFTP transfer throughput characteristics (shows variance even for mem2mem)

	Min	1st Qu.	Median	Mean	3rd Qu.	Max	CV
Oct	2196	4327	4744	4683	5065	6577	0.15
Dec	531.6	3408	4463	4202	5056	6290	0.29

Correlation coefficients with CPU usage

	NERSC CPU	SLAC CPU
Oct	0.82	0.90
Dec	0.95	0.98

Correlation coefficients with SNMP bytes on SLAC router links

	farm02a	core1	border2
Oct	0.99	0.95	0.77
Dec	0.99	-0.01	0.25

← Path changed?

Prototyping

- Our hypothesis is that we need a co-scheduling system for server and network resources to reduce throughput variance
- Suggestions for prior co-scheduling work that can be reused?

Summary

- Policing cuts throughput
  - if CPU and I/O resources are primary determinants of throughput, then do circuits hurt or help?
- Primary cause of throughput variance:
  - CPU and I/O resources of DTNs
  - Not so much the net (overprovisioned)
- Feedback/questions?
  - please email mvee@virginia.edu